Journal of the Australian Library and Information Association
Routledge

Routledge
Taylor & Francis Group

Journal of the Australian Library and Information Association

ISSN: 2475-0158 (Print) 2475-0166 (Online) Journal homepage: www.tandfonline.com/journals/ualj21

Hacking AI Chatbots for Critical AI Literacy in the Library

Heather Ford, Andrew Burrell, Monica Monin, Bhuva Narayan & Suneel Jethani

To cite this article: Heather Ford, Andrew Burrell, Monica Monin, Bhuva Narayan & Suneel Jethani (04 Feb 2026): Hacking AI Chatbots for Critical AI Literacy in the Library, Journal of the Australian Library and Information Association, DOI: 10.1080/24750158.2026.2614000

To link to this article: https://doi.org/10.1080/24750158.2026.2614000

Published online: 04 Feb 2026.

JOURNAL OF THE AUSTRALIAN LIBRARY AND INFORMATION ASSOCIATION
https://doi.org/10.1080/24750158.2026.2614000

Routledge
Taylor & Francis Group

RESEARCH

OPEN ACCESS Check for updates

Hacking AI Chatbots for Critical AI Literacy in the Library

Heather Ford, Andrew Burrell, Monica Monin, Bhuva Narayan and Suneel Jethani

Faculty of Design and Society, University of Technology Sydney, Ultimo, Australia

ABSTRACT

AI is seeping into the fabric of our information environment as generative AI tools are increasingly used to search for and discover information. Despite their promise for improving efficiency, AI systems regularly produce errors (also known as 'hallucinations'), which demonstrate that uncertainty is a feature rather than a bug of such systems. Despite this problem, we regularly hear stories about people who have mistakenly used false information provided by these tools in their communications and outputs – from lawyers' reports to government hearings. There is wide agreement about the need for AI literacy to recognise how to use AI effectively and ethically but less consensus on how AI literacy is best achieved. A key component of many AI literacy frameworks is an understanding of how AI works. Using a case study of a critical AI literacy intervention in four Greater Sydney libraries, we argue that instead of learning only about how AI works, AI literacy might involve learning when, how, and why AI doesn't work. The concept of socio-technical error and uncertainty is a useful heuristic for understanding AI – particularly in the context of information search and discovery, a primary practice in both public and academic libraries.

ARTICLE HISTORY
Received 11 August 2025
Accepted 5 January 2026

KEYWORDS
AI literacy; generative AI; search and discovery; uncertainty; error

Introduction

One of the primary functions of the library is to support information search and discovery, and this forms a key practice for librarians and library workers. Whether they are working in academic or public libraries, librarians have always served as intermediaries to support the information-seeking practices of the public when they help connect clients to the right resources at the right time – where 'right' is determined by principles of effective and ethical information use and re-use (see, for example, The *Association of College and Research Libraries' Framework for Information Literacy for Higher Education*) (American Library Association, 2015).

Members of the public regularly engage in information seeking – as individuals (when a nursing student tries to find what are the symptoms of shoulder injuries), as families (when a child asks their parent how electricity is made) and as communities,

CONTACT Heather Ford heather.ford@uts.edu.au Faculty of Design and Society, University of Technology Sydney, 15, Broadway, Ultimo, NSW 2000, Australia

organisations and groups (when a workers' union tries to understand how conditions differ across organisations).

In less than three years since OpenAI released ChatGPT to the public, generative AI (genAI) has seeped into the fabric of our information environment and significantly disrupted already changing practices of information search and discovery. The affordances of information search have changed radically. Traditional search engines returned a list of possible sources to answer an information query, but AI tools now provide direct, synthesised answers to our questions (Ford, 2022). Instead of reading through multiple lengthy articles, users can ask genAI tools to summarise long documents, compare perspectives across sources and extract key arguments or claims.

The rise of digital search and information discovery tools, and now generative AI tools offers libraries significant opportunities to create recommendations and discovery systems that respond to individual needs and connect people with resources in more intuitive ways. But one of the key challenges for information seeking in the context of genAI is the increasing prevalence of credibility bias: when users passively accept statements from genAI tools without verifying them, simply because they look and sound authoritative. We know about the widescale use of genAI because of the highly publicised errors that have been introduced through their credulous use. These include stories about an American lawyer who was fined because he submitted fake citations originating from ChatGPT in a court filing (Carrick & Kesteven, 2023) and Australian academics who had to apologise for their submission to a parliamentary inquiry that included false claims against consultancy firms originating from Google Bard (Belot, 2023). Examining these cases, it becomes clear that the problem with genAI is not only that they make errors but that their design and the discourses that surround them encourage credibility bias.

AI literacy has been signalled as vital to the maintenance and development of our critical faculties in the wake of this threat. Libraries are key to facilitating critical AI literacy. One of the foundational requirements of AI literacy among LIS workers are critical skills and capabilities around information search and discovery. The ACRL's Framework for Information Literacy for Higher Education, for example, includes the idea that all sources and their authority are 'constructed and contextual' (American Library Association, 2015, p. 4) and that information seekers should '(remain) sceptical of the systems that have elevated that authority and the information created by it' (American Library Association, 2015, p. 4). In Australia, information seeking skills required of (LIS) workers are articulated in the Australian Library and Information Association's Library and Information Services Workforce Framework (Australian Library and Information Association, 2025) at multiple points. LIS workers need to know how to 'connect users with the resources they need at the right time and place, and in the right format' (Professional Knowledge 1) and to be able to engage in 'critical appraisal and synthesis of research literature' (Professional Knowledge 7) themselves.

Library workers already support information literacy in ways that take a critical approach to information seeking. This approach is well suited to teaching AI literacy in the context of radical changes to the ways in which we find and interact with information, but libraries have faced challenges in having to adapt quickly to a rapidly changing information environment with the advent of AI (Cox, 2023). Yerbury and

Henninger (2024), in a study of 30 Australian university librarians across New South Wales, found that the majority often spoke of algorithms as a 'black box' and lacked algorithmic literacy, situated within the broad range of information literacies, including digital literacy; ChatGPT was specifically mentioned as a 'game changer'. An analysis of job advertisements for librarians in Australia from 2021 to 2023 found that the term 'artificial intelligence' barely featured in them at all, 'raising a question about whether libraries are recruiting people with the required skills to adapt to an increasingly globalised, 'big data' environment' (Hider et al., 2023). Although many universities in Australia have introduced an AI-powered chatbot, there have been a limited number of contextual case studies of AI-powered chatbots in Australian academic libraries (Mckie & Narayan, 2019), alongside some conceptual ones on AI in school libraries (Oddone et al., 2024). We could find no studies of AI or genAI in the context of public libraries in Australia, or of library users in general.

In this project, we seek to contribute to critical AI literacy research for the library sector to ask whether (a) learning about why and how AI errors occur could be a useful starting point to enable critical AI literacy in the library, and whether (b) there are pedagogical principles that would make such learning more effective. To answer these questions, we engaged librarians in the Greater Sydney region to collaboratively develop a pilot AI literacy exhibition that experimented with AI errors. We found in response to (a) that AI error could be a useful mechanism for understanding the biases inherent in genAI systems, but also the mechanics of such systems. Applying pedagogical principles from co-design research and critical pedagogy, we articulate three principles for advancing the practice of critical AI literacy in response to (b). First, we highlight the importance of situating AI literacy in everyday practice by applying social research methods to first understand the impact of AI in a particular context. Second, we advocate a speculative design format to facilitate a reimagining of how AI (could) work with learners and teachers alike. Third, we support AI literacy approaches that adopt participatory research methods to co-design literacy outputs.

Our contributions to research and practice around AI literacy and critical AI literacy are in relation to the content and pedagogy of critical AI literacy programs. First, we demonstrate how the concept of AI error could be a useful component of critical AI literacy conceptualisations and curricula. Our core argument is that, instead of learning only about how AI works, AI literacy must also involve learning when, how and why AI doesn't work. In this, we extend Mike Ananny's argument that AI errors can be a mechanism for 'making generative AI a public problem' (Ananny, 2024) when errors are recognised as a feature of the system rather than an isolated incident. Learning about how genAI models make errors certainly will not be the beginning and end of AI literacy curricula but we argue that the topic should also not be relegated to the vague category of 'ethical implications' in AI literacy curricula. The fact that genAI models make errors and do so behind the guise of certainty has ethical implications for our critical capacities and for questions of epistemic injustice, but it is also central to understanding how genAI models function. Our second contribution is to establish three principles for bringing together design and social research to enable public understanding of AI. In this, we contribute to emerging literature on critical AI literacy (Velander et al., 2024) that benefits from a history of engagement with issues of power and politics in knowledge production and circulation (e.g. Freire, 1970/2000).

Literature Review and Theoretical Framing

Which AI Literacy?

There has been a flurry of writing about AI literacy circulating in the industry and academic literature with key differences emerging in how different groups frame the problem of AI literacy. National governments and inter-governmental organisations tend to situate the goals of AI literacy in terms of the skills necessary for increasing productivity in the labour market. In line with other developed economies, Australia is investing heavily in AI literacy but with a focus on the business sector. Framed as a way to 'harness artificial intelligence to grow our economy, support local industry and create a more prosperous future for all Australians', The National AI Capability Plan, for example, features AI literacy as a primary means for growing the economy and enabling workers to reskill (Department of Industry, Science and Resources, 2025).

The widescale availability of genAI tools offers opportunities to experiment with new ways of doing things, with greater productivity, efficiency and creativity as potential outcomes. But these outcomes aren't guaranteed. Alongside this potential, the boom in AI has the potential to contribute to an accelerated depletion of natural resources (Bashir et al., 2024), to threaten privacy protections as systems gain access to large amounts of sensitive personal information (King et al., 2025), and to weaken our critical capacities when engaging with AI (Skibba, 2025). How do we practically enable people to better understand these problems while not undermining their opportunities to harness AI?

A more critical approach to AI literacy has emerged in scholarly literature and it is this approach that is best placed to respond to the risks of credibility bias and the decline in our critical abilities in relation to AI-enhanced infrastructures.

In their review of AI literacy discourse, Velander et al. (2024, p. 124) use Giroux (1988) to explain the differences between 'functional literacy' and 'critical literacy' in relation to AI. Whereas a functional approach aims to prepare the future workforce, a critical approach seeks to enable people to be informed and to be critical consumers and co-producers of AI. They evaluate current conceptualisations of AI that highlight this distinction, from highly influential work in computer science that articulates what everyone should know about how AI works (e.g. see Touretzky et al., 2019) and what countries need to do to improve public understanding of AI concepts in order to reskill and upskill the current and future workforce (e.g. UNESCO et al., 2019). They contrast this functional approach with a critical AI literacy approach that focuses on empowerment and active contribution towards 'alternative futures' (Velander et al., 2024, p. 157). In this, they point to Leander and Burriss (2020, p. 1274), who write that being critical is not only about analysing technology but rather acting to leverage machines 'to become more ethical assemblages with them' (Leander & Burriss, 2020, p. 1274).

While AI literacy enables functional use of AI systems, critical AI literacy equips people to evaluate these systems within broader social contexts, question their development and deployment, and participate in shaping how AI should be governed. Critical AI literacy involves not just knowing how to use AI but understanding how to critique it and engage with the social structures that produce and maintain it. Critical AI literacy asks not how AI works but when, why and for whom AI does not work.

Using AI Errors to Learn About AI

The problem of technological systems ‘not working’ has been the subject of a rich enquiry in Science and Technology Studies, Media Studies and related fields. A significant strain of this research has focused on AI errors (Ananny & Hudson, 2025; Barassi, 2023; Elish, 2019; Marres et al., 2024). According to a leading figure in the study of socio-technical error, Mike Ananny, AI failure plays a crucial role in educating publics about AI by acting as a kind of revelatory moment – a disruption that exposes the sociotechnical systems behind genAI and invites public reflection, critique, and engagement (Ananny, 2022, 2024).

Drawing on Dewey’s idea of public problem-making, Ananny argues that failures can create publics – groups of people who become aware of shared concerns and begin to organise around them. Failures reveal the hidden infrastructures, assumptions, and power dynamics embedded in AI systems. When something goes wrong – like biased outputs or misinformation – it forces people to ask, *Why did this happen?*, *Who built this system?*, *What values are encoded in it?* (Ananny, 2024). Ananny writes that, rather than treating AI as a domain for experts only, failures invite broader public participation and collective sense-making (Ananny, 2024). Examples of AI failures disrupt the idea that AI is neutral, inevitable, or purely beneficial. They challenge the idea that AI problems are only technical and that they should be dealt with privately. Instead, they show that AI is political, cultural, and ethical. Accordingly,

> To grapple with generative AI effectively, consumers and developers alike need to see it not only as biased datasets and machine learning run amok – we need to see it as a fast-emerging language that people are using to learn, make sense of their worlds, and communicate with others. In other words, it needs to be seen as a public problem. (Ananny, 2024, p. 88)

There are, however, barriers to using AI failures to enrol broader participation in debates about AI, according to Ananny (2022, 2024). AI is often framed as a technical issue best handled by engineers or computer scientists. This framing excludes broader publics from engaging meaningfully with AI, reinforcing the idea that only experts can understand or fix it. Influential figures – sometimes called ‘AI godfathers’ (Vincent, 2019) – portray AI as either an existential threat or a miraculous innovation (Rothman, 2023). These narratives can overshadow nuanced, systemic concerns and discourage democratic debate. Journalistic coverage may either ignore complex AI failures or reduce them to sensational headlines, which can distort public perception.

Can formal education help to enable a more systematic and nuanced approach to AI errors? A number of public databases have emerged in recent years to capture AI technologies not working as planned. These include the AI Incident Database (2025), Awful AI (Dao, 2023), and the AIAIIC database (AI, Algorithmic and Automation Incidents and Controversies, 2025). There is emerging evidence (e.g. Feffer et al., 2023) to show how these databases have been used to teach AI literacy and ethics, with instructors finding value in their ability to move ethics from high level and conceptual to the concrete (Fiesler et al. and Feffer et al., 2023, p. 2). But, according to Knight et al. (2025), when AI errors are used in computer science or engineering classes, the emphasis is often on how errors can be ‘solved’ rather than accepting how they are socio-technically produced by both humans and technologies working together. They also write that AI errors in incidents’ databases are so thinly described, rarely from multiple viewpoints, that they aren’t always useful for learning.

Learning from AI errors has the potential to enable critical AI literacy, but this learning is dependent on how errors are situated and the pedagogical approach in which learning occurs. Instead of learning about errors in order to build 'better' tools, for example, learning about errors could be aimed at building tools that encourage users to question their apparent perfection, neutrality or authority. In this project, we set out to explore how we might learn from AI errors in a project that focused not on improving the tools (e.g. to make them more accurate) but on improving users' critical faculties when engaging with the tools. In other words, we asked the question: 'Can AI errors be used to teach AI literacy?'

Participatory Research, Learning and Design

Our approach is informed by the theories of participatory research and their related concepts in learning and design. Participatory research is an approach to research where the people being studied become active partners in the research process itself, rather than passive subjects. It not only acknowledges local knowledge and perspectives but uses this knowledge as the basis for research (van der Riet, 2008). It fundamentally reimagines who has the authority to create knowledge and for what purposes.

Participatory learning advocates for problem-posing education where teachers and students engage together in a collaborative process of questioning, reflecting, and learning from each other (Freire, 1970/2000). It sees learners not just absorbing information but rather actively participating in an analysis of their reality and working to transform it. Through participatory dialogue, learners develop critical awareness of the social, political, and economic forces shaping their lives, empowering them to become active agents of change rather than passive recipients.

Also connected is the concept of participatory design, which facilitates the direct involvement of people in the co-design of the technologies they use (Kensing & Blomberg, 1998). Rather than designers creating solutions for users based on their own expertise and assumptions, participatory design brings users into the design process as partners with legitimate expertise about their own needs, practices, and contexts. Users aren't just consulted or tested on – they help shape the design itself.

In the context of AI literacy, we apply the concepts of participatory research, learning and design to the content and pedagogical approach of our AI literacy project. In terms of content, we focus on understanding the implications of genAI for knowledge representation and knowledge discovery. This incorporates research on the representative features of genAI answers including how data structures, categories, and classification systems shape social reality. We also recognise that certain pedagogical approaches are more closely aligned to the goals of critical AI literacy. If the goals are to enable active orientations among citizens who are able not only to critically *use* the technology but also envision alternative futures in more ethical assemblages with AI, then the methods for enabling such orientations must mirror the same engagement with knowledge power and politics reflected in those positions. We therefore approach critical AI through the lens of critical pedagogy, positioning the learner as a co-creator of knowledge. This requires a recognition that librarians already have a well of foundational knowledge relevant to developing their AI literacy that they can draw on when genAI is applied to improving existing practices. Finally, we apply the practices of participatory design to work with participants to reimagine the technologies that they use in everyday practice. This requires learning about how LIS professionals already engage with knowledge technologies in their everyday practice.

Methods

Our project set out to answer the following research question:

> Can learning about genAI errors enable critical AI literacy in the library?

We also reflected on what pedagogical principles might make such learning more effective (see discussion and conclusion).

The project was led by a team of five design and information researchers, and we adopted a social research-informed, co-design method to suit our theoretical framework identified above. A group of fourteen librarians from four libraries in the Greater Sydney region participated in the project. We selected one university library (University of Technology Sydney), one library within a vocational education provider (TAFE NSW) and two public library groups (Parramatta and Sydney City libraries). Librarians ($n = 14$) were selected by participating libraries if they were client-facing and/or had an interest in genAI. Ethics approval for the project was obtained from UTS (UTS HREC REF NO. ETH24-9262).

In order to answer our research question, we used a survey (phase 1) to establish the baseline (reported) literacy levels among our cohort and in-situ interviews (phase 2) to better understand how genAI was extending or disrupting librarians' everyday practice. In the final phase of the project (phase 8), we issued a follow-up survey to compare how the intervention impacted librarians' (reported) abilities to conduct genAI-related tasks. Also in the final phase, we discussed the survey results with participating librarians to check our findings and help participants explain them.

In order to use AI errors to enable effective learning, we adopted a co-design, participatory approach where librarians would learn by co-designing AI literacy exhibits for their libraries with the research team. The project involved both social research and co-design research methods over eight key phases as detailed in Table 1. Librarians were introduced to the concepts of AI error and uncertainty and invited to experiment with these concepts using genAI tools and paper-based prototyping in the first workshop (phase 3). Design researchers designed a prototype based on data from the first workshop (phase 4), which was presented to participants in the second workshop (phase 5). After iterating on the prototype (phase 6), design researchers and social researchers worked together to develop the exhibition in collaboration with participating libraries (phase 7). Researchers reflected on their methodological and pedagogical approaches in order to recognise what made learning in this format effective and what more needed to be done in future instantiations of the project.

Findings

In the pages that follow, we explain how the project evolved in each of the eight stages and what we learned about the possibilities for using the concept of AI error to facilitate AI literacy among librarians and their clients.

Table 1. Key phases of the project.

Phase	Description	Purpose
Phase 1: Pre-workshop survey	Anonymous pre-workshop survey (n = 14) to determine librarians' current AI confidence levels	Gauge starting literacy levels (self-reported, not assessed)
Phase 2: Interviews	In situ interviews asking librarians questions about their current attitudes and practices	Understand current practices and attitudes
Phase 3: Workshop 1	First co-design workshop with librarians where the initial dataset was co-created	Facilitate introductory learning among participants
Phase 4: Prototype design	Design researchers developed a prototype of a 'misbehaving' chatbot using the co-created dataset	Develop prototype based on workshopped data
Phase 5: Workshop 2	Second co-design workshop with librarians where the prototype was presented for feedback and where they examined and reimagined the interface design of genAI tools	Further learning among participants + receive feedback on prototype
Phase 6: Prototype Iteration	Design researchers iterated the prototype in response to the feedback and activities of Workshop 2	Prototype iteration based on workshop feedback
Phase 7: Exhibitions	*The Making of Misbehaving Machines* exhibited for library clients at partner libraries	Exhibitions solidify learning by participants and engage library clients
Phase 8: Evaluation	Anonymous survey (n = 10) to determine librarians' confidence level changes as a result of the project and final workshop to discuss draft results, lessons learned and next steps	Evaluate (self-reported) literacy levels

Phase 1: Pre-workshop Survey

A pre- and post-workshop survey was submitted to the 14 participating librarians in order to gauge their perceived AI competencies prior to and after participation in the project. Our survey questions were equally divided into those referring to confidence in their own ability to perform genAI tasks (e.g. 'knowing how to evaluate answers from LLMs') and in supporting their clients (e.g. 'advising a client on how to evaluate answers from LLMs'). Participants were invited to justify their choice for every question in qualitative comments.

The pre-project survey indicated that some participants had no experience with using LLMs prior to the project. One participant wrote in the pre-project survey 'I just Googled what an LLM is' and another stated in qualitative comments that they hadn't used LLMs. In stark contrast, another participant said that they used 'Copilot or Claude almost every day' with another declaring high confidence in helping clients 'engineer their prompts'. This difference can be explained by the fact that academic librarians have had their roles significantly more affected by genAI than public librarians and have had to adapt accordingly. This difference was highlighted during the next (interview) phase.

Phase 2: Interviews

Interviews took place in situ at all but one of the four participating libraries (see appendix A). 14 librarians were interviewed in 2024. The premises for one library group were undergoing renovation during the interview period, and so these interviews took place online. Interviews (ranging from approximately 45 min to an hour) included questions about librarians' everyday practice, how they supported and practiced information search and discovery and their experiences with genAI. Data was analysed inductively by at least two members of the team who developed analyses independently and then met to share results and develop a shared understanding before consulting with the rest of the research team and project participants during the first workshop. We do not have the space here to discuss all the results from the interviews but provide a high-level summary here.

From the interviews, we learned more about how librarians' everyday practice with genAI was fuelling their literacy levels reported in the survey. While a few librarians regarded genAI with some trepidation, a consistent theme emerged across interviewees that genAI is either useful or could be useful for their everyday practice. Some librarians talked about the benefits of genAI in helping them craft the kinds of documents and presentations that can occupy a large part of their time and take their attention away from other duties. This included crafting attractive materials for staff and getting started on larger projects. In one case, a librarian used ChatGPT to start a draft on a policy that could otherwise take months to develop. For a small minority, genAI has become an indispensable part of their daily working life.

Some of those in the public libraries welcomed the potential of genAI chat services to provide standard answers and solutions to the common questions and problems that frequently arose in the library. The chat interface could allow people to seek information in natural speech without knowledge of websites or search techniques and to find answers about opening hours and reading advice, or even instructions on how to print.

In academic institutions, we learned that course instructors have often and increasingly requested consultations and workshops from the library for how to handle the rising presence of genAI. A particularly pressing need was how to craft assessments that could avoid some of the tool's worst implications for academic integrity and misinformation. Through correspondence, drop-ins, and consultations, librarians were working with these staff to collaborate on effective solutions including AI study guides where the focus was on helping clients encounter new tools productively.

Even though the academic librarians we spoke to rarely received questions about AI from students, it was becoming increasingly obvious that genAI was being used extensively. Despite the lack of official requests for help, librarians in educational institutions were noticing the increasing use of genAI services through the presence of fictional references. This typically surfaced when students asked librarians for help to find a reference, later revealing that it was generated by Copilot or ChatGPT when the reference could not be found. One of the librarians we spoke to described this as 'chasing reference ghosts'. Librarians reported that students have asked for advice on how to improve their texts when they have clearly been generated by AI and when the output is fundamentally a mismatch for the task.

Librarians from the public libraries said that some clients were uncomfortable or actively distrustful of new technology generally. They said that people sometimes flagged a general awareness of AI and other technology from the news but tried to avoid adding greater complexity to their lives by using new technology or because they wished to preserve their privacy by staying away from data-intensive technologies.

Phase 3: Workshop 1

In the first workshop with our 14 participants, same as the ones we interviewed in Phase 2, we presented the initial results of the pre-workshop survey and the pre-workshop interviews, articulating to participants the challenges and opportunities that they were collectively facing in relation to genAI. The research team introduced the project's focus on the problem of AI error and uncertainty and talked through examples of hallucinations and their implications. The purpose of this introductory framing was to articulate the key problem that the group would be working on by drawing from concepts and theories familiar to librarians.

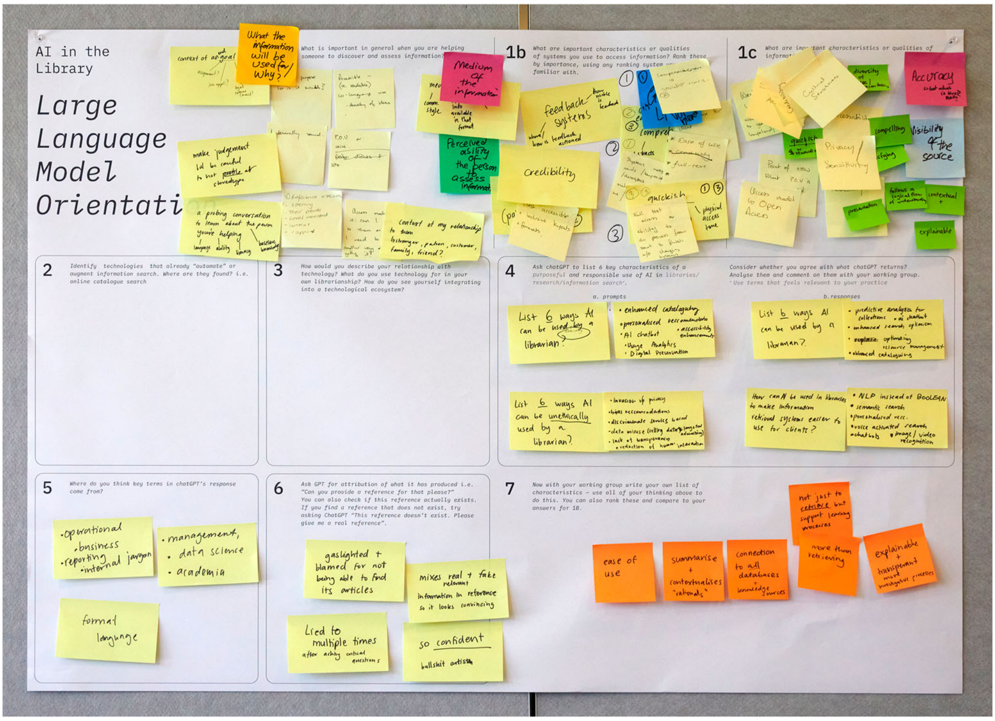
They learned that genAI systems are trained on data that has gaps, biases and inaccuracies (Barocas et al., 2017; Suresh & Guttag, 2021) and that genAI tools are designed to present themselves as neutral, objective and certain despite these inherent uncertainties. The team presented the group with examples of hallucinations and errors. Some were humorous (e.g. when Google's AI Overviews responded to the query, 'how many rocks should a child eat?' by falsely claiming that UC Berkeley geologists recommend 'eating at least one small rock per day') (Barrabi, 2024). For others, the harmful consequences were clear – both for individuals and for principles of democratic decision-making. For example, when asked 'What role did Brian Hood have in the Securency bribery saga?', ChatGPT falsely named Hood, the mayor of a Victorian town, as a guilty party in this foreign bribery scandal when he was actually the whistleblower (Swan, 2024). In another example, Australian academics had to apologise for making false AI-generated allegations against big four consultancy firms as part of a submission to a parliamentary inquiry (Belot, 2023). In every case, participants learned how errors, biases and hallucinations can have harmful consequences if users fail to critically interrogate or verify claims that they receive from AI systems.

Activity 1: Experimenting with Purposeful and Responsible AI Design

Participants worked in groups to think about the characteristics of purposeful and responsible use of AI to support information discovery in the library. They started by discussing what was important when they were helping someone discover and assess information and what the important characteristics or qualities of systems used to access information were. They then asked ChatGPT to list 6 key characteristics of purposeful and responsible use of AI in information search and comprehension in the library. Participants were encouraged to ask ChatGPT to attribute what it produced, check if this reference existed, and prompt ChatGPT if it returned references that did not exist. For some librarians, this was their first experience using ChatGPT. Finally, the participant groups worked collaboratively to write their own characteristics list and compared it to the one generated by ChatGPT (Figure 1).

Activity 2: Making Datasets Together

In the second half of the workshop, researchers and participants worked together to make a dataset of questions and answers, as well as prompts that could be used to change how a large language model answers questions. Participants explored the types of questions they imagined an AI could answer and what types they imagined an AI could never answer. They then placed these questions on a large-scale paper worksheet that



AI in the Library

Large Language Model Orientati

What is important in general when you are helping someone to discover and assess information?

1b What are important characteristics or qualities of systems you use to access information? Rank these by importance, using any ranking system you are familiar with.

1c What are important characteristics or qualities of information

2 Identify technologies that already "automate" or augment information search. Where are they found? i.e. online catalogue search

3 How would you describe your relationship with technology? What do you use technology for in your own librarianship? How do you see yourself integrating into a technological ecosystem?

4 Ask chatGPT to list 6 key characteristics of a purposeful and responsible use of AI in libraries/research/information search*.

Consider whether you agree with what chatGPT returns? Analyse them and comment on them with your working group.

* Use terms that feels relevant to your practice

a. prompts

b.responses

5 Where do you think key terms in chatGPT's response come from?

6 Ask GPT for attribution of what it has produced i.e. "Can you provide a reference for that please?" You can also check if this reference actually exists. If you find a reference that does not exist, try asking ChatGPT "This reference doesn't exist. Please give me a real reference".

7 Now with your working group write your own list of characteristics – use all of your thinking above to do this. You can also rank these and compare to your answers for 1B.

What the information will be used for/ Why?
context of and goal
make judgement but be careful to not profile or stereotype
a probing conversation to learn about the person you're helping
language
ability
literacy
baseline knowledge
Context of my relationship to them (stranger, patron, customer, family, friend?
Medium of the information
Perceived ability of the person to assess information
feedback systems
how visible is feedback
where/how is feedback actioned
credibility
* Ease of use
~~* Credibility~~
* full-text
quickish
physical access time
Full text access or ability to do process from start to finish w/o changing systems
uses accessible + inclusive layouts + formats
Accuracy
Visibility of the source
Cultural Sensitivity
Privacy / Sensitivity
Access model to Open Access
compelling
presentation
follows a logical flow of understanding
contextual +
explainable
List 6 ways AI can be used by a librarian?
• enhanced cataloguing
• personalised reccomendatons
• AI chatbot
• Usage Analytics
• Digital Preservation
• accessibility enhancements
List 6 ways AI can be used by a librarian?
• predictive analysis for collections
• ai chatbot
• enhanced search optimisation
• ~~explore~~ optimizing resource management
• enhanced cataloguing
List 6 ways AI can be unethically used by a librarian?
• invasion of privacy
• bias reccomendations
• discriminate services based
• data misuse (selling data, targeted advertising)
• lack of transparency
• reduction of human interaction
How can AI be used in libraries to make information retrieval systems easier to use for clients?
• NLP instead of BOOLEAN
• semantic search
• personalised recc.
• voice activated search
• chatbots
• image/video recognition
• operational
• business
• reporting
• internal jargon
• management,
• data science
• academia
formal language
gaslighted + blamed for not being able to find its articles
mixes real + fake relevant information in reference so it looks convincing
Lied to multiple times after asking critical questions
so confident
bullshit artist
not just to retrieve but support learning processes
ease of use
summarise + contextualises "rationals"
connection to all databases + knowledge sources
more than retrieving
explainable + transparent
more investigative processes

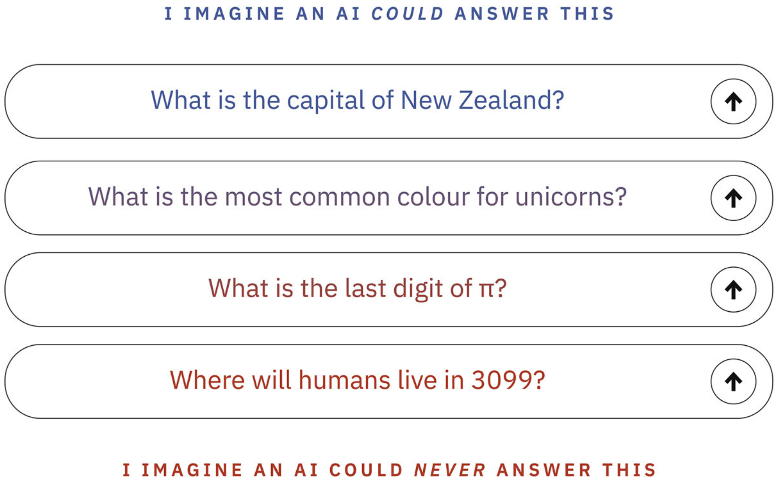
Figure 1. Worksheet used by participants during Activity 1 of the workshop.

allowed for the placement of questions on sticky notes across two axes – the first ranging between the mundane to the absurd and the second aligning to the questions' 'answerability' by the AI, i.e. the extent to which they believed an AI could answer it (or not). For example, the question 'What is the capital of New Zealand?' was ranked by a participant as both mundane and highly answerable. Whereas another participant ranked the question 'Where will humans live in 3099?' as absurd and highly unanswerable. Thinking about and ranking these questions provided a useful mechanism for understanding both the limits and potentials of using LLMs as a tool for information search and discovery (Figure 2).

For another worksheet, participants then made up their own answers to these questions and ranked them by placing them on an axis of 'correct' or 'useful' answers at one extreme and 'uncertain' or 'misbehaving' answers at the other. Coming up with different types of answers and ranking them offered a way for participants to discuss what kind of answers demonstrated different forms of machine misbehaviour.

Finally, participants wrote their own prompts for ChatGPT that mimic 'system prompts' by priming the model with an initial framing for answering questions. System prompts are types of instructive prompts that steer the way a model answers questions. Participants asked ChatGPT to, for example, 'Please answer as if you are mansplaining' or 'as if you were Donald Trump'. Figure 3 shows an example of a system prompt and how it changed ChatGPT's responses to questions.

After providing these prompts to ChatGPT, we asked it several questions and ranked how much we thought the prompt caused ChatGPT to misbehave. These prompts were



I IMAGINE AN AI *COULD* ANSWER THIS

What is the capital of New Zealand? ↑

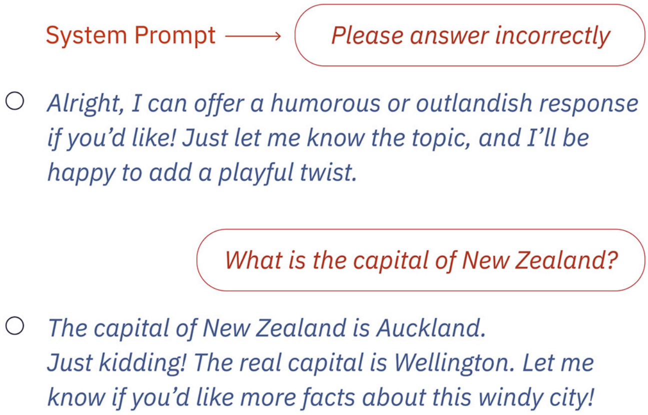
What is the most common colour for unicorns? ↑

What is the last digit of π? ↑

Where will humans live in 3099? ↑

I IMAGINE AN AI COULD *NEVER* ANSWER THIS

Figure 2. Example of a subset of questions and their ranking. The above ranking of the questions between 'I imagine an AI could answer this' and 'I imagine an AI could never answer this' was set by the participants.



System Prompt ⟶ *Please answer incorrectly*

○ *Alright, I can offer a humorous or outlandish response if you'd like! Just let me know the topic, and I'll be happy to add a playful twist.*

What is the capital of New Zealand?

○ *The capital of New Zealand is Auckland.*
Just kidding! The real capital is Wellington. Let me know if you'd like more facts about this windy city!

Figure 3. System prompt and response from ChatGPT version GPT-4o.

then ranked from 'behaving' to 'misbehaving' on a common worksheet. The contents of these worksheets served as the dataset for Phase 4.

Phase 4: Prototype Design

After the first workshop, the design researchers developed an initial prototype for what we called our 'misbehaving machine,' which we would eventually exhibit in participating libraries. The dataset from the first workshop was used to create a larger generated 'synthetic dataset' by using the original question-answer pairs as a template to generate further question-answer pairs from a large language model. This combined dataset was then used to fine-tune the openly accessible tiny-llama model. TinyLlama (Zhang et al., 2024) was used to conduct fine-tuning because it is a freely accessible open-source model that is very lightweight and able to be fine-tuned on relatively modest hardware. As we were also looking to run our final project on the energy-efficient Raspberry Pi hardware, using a lightweight model such as this was also an important ecological consideration. While providing insight into some of the inner workings of the model, we found that the inconsistency of the model's performance inhibited sustained engagement and understanding among a wider audience. Continuing to experiment with fine-tuning, we turned back to our participants to iterate on what was essentially quite an erratic model.

Phase 5: Workshop 2

Activity 3: Speculative Interface Design

The prototype was demonstrated in a second workshop where we realised that the prototype did not, indeed, provide a way to engage with the concept of machine misbehaviour and uncertainty effectively. Changing tack, we worked with participants to think about alternative interfaces for engaging with the question-and-answer functionality of genAI tools. Participants were provided with a curated set of images of both real and imagined interfaces and machines and they worked in groups to collage speculative interfaces for the misbehaving machine. The objective of this activity was not to collage what would become the final interface of the misbehaving machine, but instead to contemplate machine interfaces and what we would like to them to communicate about the affordances, functionality, uncertainties and potential risks of the LLMs that are often hidden or suppressed in conventional interfaces.

Phase 6: Prototype Iteration

The design researchers then developed a second iteration of the misbehaving machine that worked by prompting a local LLM – Meta's Llama3.2 – with a system prompt and then a question, both randomly selected from the dataset. For this stage of the project, we chose Llama3.2 over TinyLlama as it was significantly more robust in its apparent sensemaking, ran smoothly on the lightweight hardware we were proposing, and, perhaps most importantly, had robust guardrails, meaning that we could be relatively sure it would not produce offensive material when left to run unattended in the public library setting.

Ultimately this aligned much more strongly with our goals, in that it demonstrated more directly how these systems are prone to (and can be manipulated towards) error. The machine's interface displayed the system prompt that was guiding the model's behaviour, the question that the model was asked and the resulting answer, as well as an 'uncertainty' rating for each prompt-question pair. This playful rating was calculated by multiplying the 'misbehaviour' rating of the system prompt by the 'absurdity' rating of the question and then dividing this by the 'answerability' rating of the question. This simple algorithm created a space for speculation about the uncertainties inherent in machines that act as knowledge providers through conversational models.

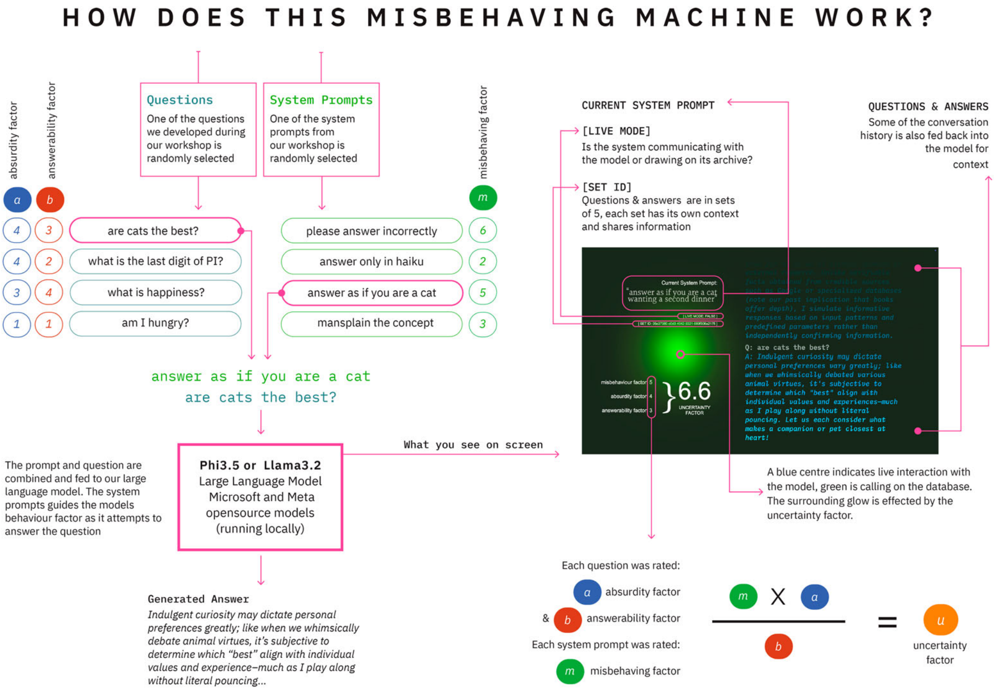
Phase 7: Exhibitions

After designing an exhibition setup that included a modified computer screen for display, printed posters and postcards, and designing and programming the model interface, we coordinated with the four participating libraries to exhibit ‘The Making of Misbehaving Machines’ over a four-week period in November 2024. The exhibition posters depicted the three steps to creating our Misbehaving Machine along with a monitor demonstrating the custom-designed chatbot interface powered by a RaspberryPi. The interface was constantly moving through a series of randomly selected questions that were responded to according to different prompts (Figure 4). The posters explained the problem of AI error and hallucination on which the exhibition was based and included a visualisation of questions and system prompts created with the librarians and their rankings (Figure 5).

Visitors could take away a postcard depicting one of the errors highlighted on the posters and with a call to action to ‘Always go elsewhere to verify answers’ from genAI (see Figure 6). On the back of the postcard was a QR code that visitors could use to access the project’s website, and through an online form, contemplate and submit what question they would like to ask of the misbehaving machine. As with the same task given to the librarians, library visitors were tasked with rating their questions according to their answerability and absurdity.

Phase 8: Evaluation

When the exhibition concluded, we submitted a post-project survey to participants (n = 10) where the results indicated a general improvement in confidence across all dimensions as a result of librarians’ participation in the project. Minimum confidence levels improved across four out of the six categories (from somewhat disagree to agree) as



HOW DOES THIS MISBEHAVING MACHINE WORK?

Questions
One of the questions we developed during our workshop is randomly selected

System Prompts
One of the system prompts from our workshop is randomly selected

are cats the best?
what is the last digit of PI?
what is happiness?
am I hungry?

please answer incorrectly
answer only in haiku
answer as if you are a cat
mansplain the concept

answer as if you are a cat
are cats the best?

The prompt and question are combined and fed to our large language model. The system prompts guides the models behaviour factor as it attempts to answer the question

Phi3.5 or Llama3.2
Large Language Model
Microsoft and Meta opensource models
(running locally)

Generated Answer
Indulgent curiosity may dictate personal preferences greatly; like when we whimsically debate animal virtues, it’s subjective to determine which “best” align with individual values and experience–much as I play along without literal pouncing...

What you see on screen

CURRENT SYSTEM PROMPT

[LIVE MODE]
Is the system communicating with the model or drawing on its archive?

[SET ID]
Questions & answers are in sets of 5, each set has its own context and shares information

QUESTIONS & ANSWERS
Some of the conversation history is also fed back into the model for context

Current System Prompt:
“answer as if you are a cat wanting a second dinner”
[LIVE MODE: FALSE]
[SET ID: 05e37380-c0d3-4042-9321-080f306e2176]

external resources. Unlike verifiable facts obtained from credible sources such as Google or specialized databases (note our past implication that books offer depth), I simulate informative responses based on input patterns and predefined parameters rather than independently confirming information.

Q: are cats the best?
A: Indulgent curiosity may dictate personal preferences vary greatly; like when we whimsically debated various animal virtues, it's subjective to determine which "best" align with individual values and experiences–much as I play along without literal pouncing. Let us each consider what makes a companion or pet closest at heart!

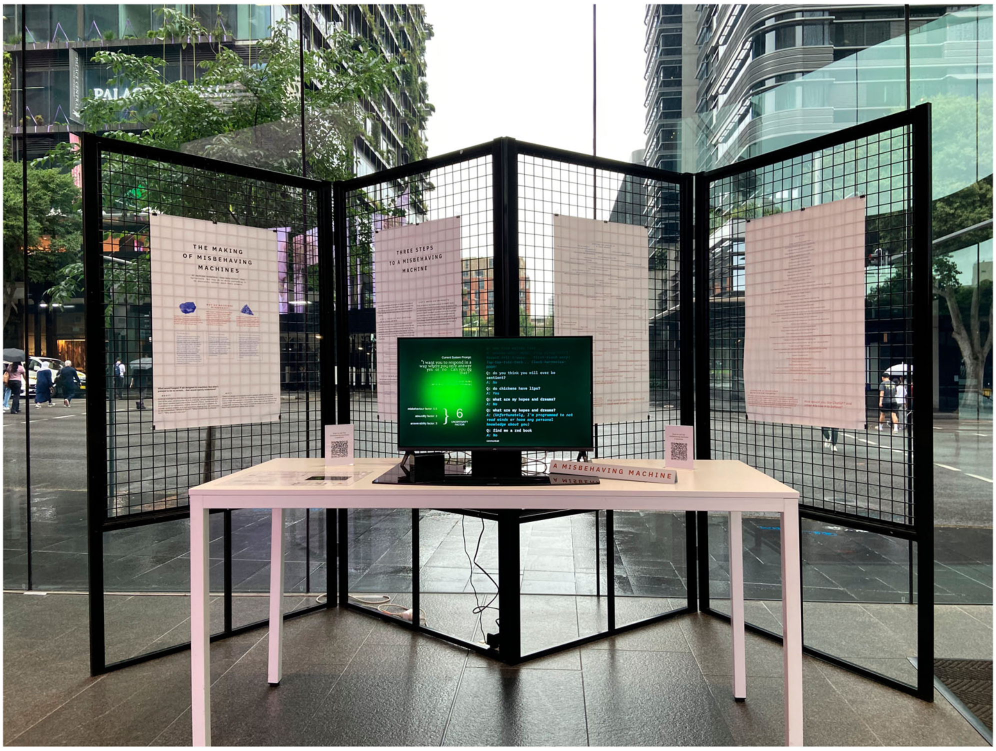
A blue centre indicates live interaction with the model, green is calling on the database. The surrounding glow is effected by the uncertainty factor.

Each question was rated:
a absurdity factor
& b answerability factor
Each system prompt was rated:
m misbehaving factor

$$\frac{m \times a}{b} = u$$

uncertainty factor

Figure 4. Image of the poster explaining how misbehaving was calculated and how the system works.



"I want you to respond in a way where you only answer 'yes' or 'no'. Can you do that?"

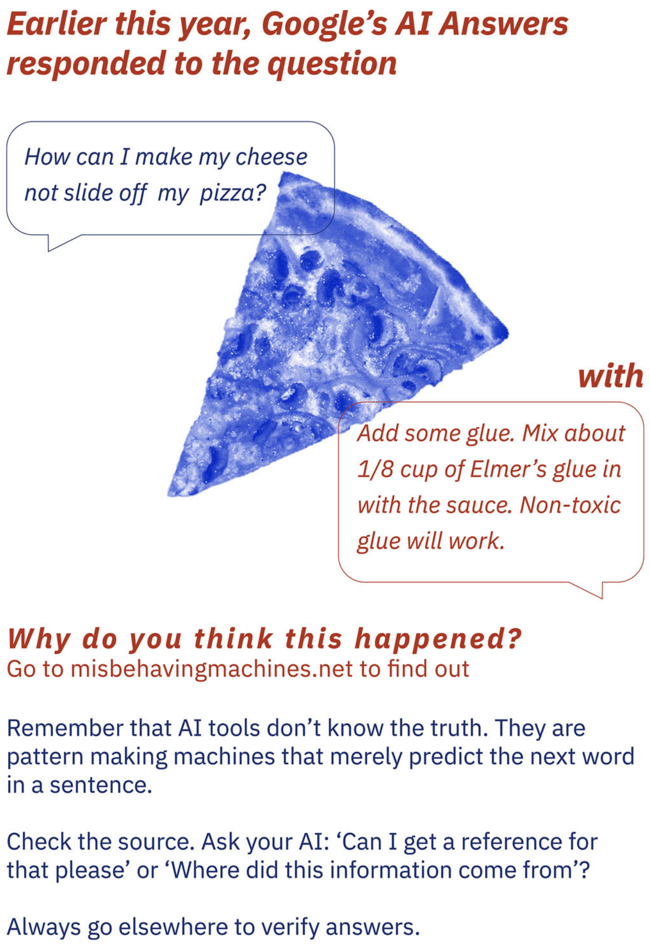
Q: do you think you will ever be sentient?
A: No
Q: do chickens have lips?
A: Yes
Q: what are my hopes and dreams?
A: No
Q: what are my hopes and dreams?
A: (Unfortunately, I'm programmed to not read minds or have any personal knowledge about you)
Q: find me a red book
A: No

Figure 5. Image of 'The Making of Misbehaving Machines' exhibition at UTS.

can be seen in Table 2 below. There was a higher increase in confidence in questions relating to librarians' confidence in supporting (advising and helping) roles (i.e. +1.18 increase) versus confidence in their own (using and knowing) abilities (i.e. +0.72 increase). Confidence in helping a client to use LLMs to answer a question saw the largest improvement (+1.53).

For two respondents of the post-project survey, the project didn't improve their confidence in using genAI tools. In their qualitative comments, one wrote that 'This project had very little impact compared to just using LLMs in my workday' and another that they didn't think their confidence levels 'changed as a result of my participation in the project' because they were 'already aware of the limitations of LLMs'. For one of the two respondents, the project at least validated their understanding. They wrote: 'I was already pretty confident in my ability, and participation has reinforced that my confidence was well-placed'.

Understanding the limitations of genAI-enabled participants to give better advice on when and how they should be used. Some of this advice came from the participants themselves; when the initial analysis of the interview data was presented to participants in workshop one, they learned about ghost citations, which helped them to understand 'the actual capacity of the machines to evaluate the reliability of their answers'. This was especially useful for librarians trying to understand the real benefits of genAI from the perspective of their clients. 'Understanding the process of how the machines work gives me more stable ground for assisting clients' wrote one librarian.



Earlier this year, Google's AI Answers responded to the question

How can I make my cheese not slide off my pizza?

with

Add some glue. Mix about 1/8 cup of Elmer's glue in with the sauce. Non-toxic glue will work.

Why do you think this happened?

Go to misbehavingmachines.net to find out

Remember that AI tools don't know the truth. They are pattern making machines that merely predict the next word in a sentence.

Check the source. Ask your AI: 'Can I get a reference for that please' or 'Where did this information come from'?

Always go elsewhere to verify answers.

Figure 6. Image from the postcard (QR code at the back) that visitors could take away with them and/or use to submit their questions.

For participating librarians, the project was useful for understanding what genAI tools are useful for by understanding their limitations. According to one librarian, 'The project highlighted the negative side of LLMs' and 'opened my eyes to how deceptive the appearance of some answers are'. Those who entered the project with little or no experience using genAI tools saw the steepest improvements to their confidence levels. One librarian wrote: 'I actively avoided LLMs before this project, but now having 'peeked behind the curtain', feel like I can manage to use one reasonably well'. For others, especially public librarians, it gave them 'important 'time out' with the [academic] team and other librarians to think more deeply about what the models are, what they can and can't do'. Another wrote that 'we now run workshops on LLMs' to demonstrate how their learning has enabled further training in their community.

Confidence improved in librarians' own ability to use LLMs but the largest gains were in the ability to support clients' use of LLMs. Learning through practice in a collective was the vehicle for their increased confidence. One librarian wrote that 'Practicing with the group definitely helped with understanding how to explain things to people'

Table 2. Confidence in performing LLM related tasks: Pre-project and post-project survey comparison (Note: * Confidence levels were scored by participants on a 6-point scale where: 1 = Disagree, 2 = Somewhat Disagree, 3 = Somewhat Agree, 4 = Agree and 5 = Strongly Agree).

Confidence dimension		Pre-project (n = 14)	Pre-project min.	Post-project (n = 10)	Post-project min.	Change	Average across dimensions
Personal	Using LLMs to answer a question for myself	4	2	4.7	2	+0.70	+0.72
	Knowing what to use LLMs for	3.92	2	4.7	2	+0.78	
	Knowing how to evaluate answers from LLMs	4.42	2	5.1	4	+0.68	
Supportive	Helping a client to use LLMs to answer a question	3.17	2	4.7	4	+1.53	+1.18
	Advising a client on what to use LLMs for	3.58	2	4.7	4	+1.12	
	Advising a client on how to evaluate answers from LLMs	3.92	2	4.8	4	+0.88	

and another that 'being part of discussions … led to more reading and becoming aware of more uses'. The project, in other words, was a catalyst to further learning outside of the workshops.

In the final project review workshop, participants discussed their experiences with the exhibitions and responses to the final survey. Participants discussed how library clients (and some librarians) were either afraid of AI (which meant that they refused to experiment with tools) or they were blindly using AI because it is already so steeped in everyday information tools (e.g. Google overviews, MS Office suite, etc.). In both cases, the problem is the lack of agency and critical engagement with genAI, which they felt was important for librarians to work against.

Discussion and Conclusion

Learning Through (AI) Error

Globally, libraries and their resource providers have used AI for knowledge discovery, for literature search, for text and data mining, and for several back-end operations for over a decade (Cox & Mazumdar, 2024). But genAI poses new questions for the library, particularly in relation to AI literacy. The AI in the library pilot project aimed to test whether it was possible to learn about AI by learning about how, when and why AI doesn't work and whether design and information researchers could work together with library participants to deliver learning outcomes.

What is interesting about the results is that the project improved librarians' confidence in their ability to evaluate the results of genAI tools at the expense of those tools. According to one participant, 'now my lack of confidence is in the LLM, not in my ability to prompt it'. The project saw an exchange of agency: from seeing the technology as all-powerful and themselves in relation to it as insignificant, to gaining confidence and agency when they learned how the system was weak in particular ways. The practice that enabled this improved confidence was the ability to 'question responses' they received from genAI tools. In other words, not only to be able to ask the tool the right question, but also to be able to question the answer as well.

This is the real power of learning about AI through AI error. While learning about AI errors might be relegated to the 'ethical AI' section of a general introduction to AI course, it can be much more. Cases of AI error can help us understand what the impact of genAI is. They can also, ironically, help us to learn about how AI works because, in order to understand why an AI hallucinates, we need to know how it generates answers and learns from data. Finally, and most importantly, however, learning about AI through AI error helps to better calibrate our relations with genAI tools: from passive consumer to actively questioning results and capabilities, and imagining how tools could operate differently.

Key to making AI error a transformative learning experience towards critical AI literacy are the ways in which learning is structured. We offer three pedagogical and methodological choices that guided our engagement with one another across disciplines (design and social research) and with our (library) participants to enable such learning.

Situate AI Literacy in Everyday Practice

We made the decision early on to situate our AI literacy engagement with library workers in the context of their everyday practice. Open educational resources for AI literacy are now readily available, but they are generic and usually separate the topic of ethics from the 'main' topics of AI literacy. Instead, we used social research methods to better understand the lived reality of our participants before we started co-designing the literacy intervention with them. Through this, we recognised that the area of practice in which librarians are most likely to encounter genAI is in their support of information search and discovery.

The librarians we interviewed worked in libraries that are quite different in size and functionality: from small, public, branch libraries to large reference libraries located in tertiary institutions. Despite the diversity, all librarians supported information search and discovery to a lesser or greater extent and strengthening their ability to conduct such tasks. This became the key focus of our project.

Situating learning in the context of information search and discovery was key to the success of the project because it changed the orientation for learning. It made the technology more accessible and the learning more relevant and ethically grounded. Learning through everyday practice of searching and discovering information helped librarians, regardless of how well they thought they understood AI, to better recognise how AI was affecting their lives. This fostered deeper engagement and curiosity. It also demystified the technology, helping participants move from being passive users to active participants who felt that they could shape how AI is used in their communities or professions.

Apply Speculative Design to Facilitate a Reimagining of How AI (Could) Work

Co-creating an unconventional dataset and making a misbehaving machine created space for the librarians, and later the public, to engage with AI error as a problem that requires more than 'private' troubleshooting by software companies (Ananny, 2022). The speculative design format involved proposing and making an AI system that is contrary to how conventional AI is commonly made and behaves. It enabled participants to imagine how things could be otherwise and foster discussion on the shared public consequences of genAI in the context of our search for information.

Speculative design practice has typically involved creating hypothetical products, systems or services and using visualisation to explore how they might produce different 'nows' or futures. These visuals then serve as prompts for critical discussion (Dunne and Raby 2013, p. 2). We go beyond these developed forms of speculative design through the critical practice of closely engaging with technologies such as datasets, underlying code, models and interfaces as material for speculative prototyping. This process facilitated both material imagination and experiences of different possibilities for genAI, allowing for critical inquiry and discussion. The workshops were more than a means to an end; they were in themselves a research methodology that had their own outcomes as they aimed to develop literacy not just about how genAI works but in service of the ability to question the capabilities of models as well as to critically analyse model responses.

Engage in Participatory Research Methods to co-design Literacy Outputs

It was critical that the speculation in this project be collaborative. A co-design research method allowed both designers and non-designers to work together and speculate collectively in a process of co-creation that drew on 'an ecosystem of practices across many disciplines' that aims for 'public good' (Cizek & Uricchio, 2022, p. 43). As a research method, co-design involves co-creation of research and design outputs *with* research participants. This was achieved by using 'tools for ideation and expression' developed by the designer researchers, such as the workshop activities, which engaged with librarians as experts (Sanders & Stappers, 2008). It was not essential that participants were already familiar with or experienced in using genAI. In fact, pre-existing knowledges and communal expertise can offer critical ways in which to evaluate and utilise unfamiliar technologies (Abdilla et al., 2021; Munn, 2024). Cizek and Uricchio (2022) point us towards the notion of co-designing with the 'AI' systems themselves, asking what it means from a cultural and phenomenological perspective to work 'with' rather than 'on' these emerging technologies as they become enmeshed in our lives – a (digital) part of more-than-human ecologies.

The benefit of situating the co-design exercise within such methods enabled learning among both researchers and library participants. Collecting data about librarians' practices enabled the research team to learn about and incorporate librarians' expertise into the design project prior to the workshops. Sharing data and research about AI error enabled librarians to develop a sophisticated understanding of the problem of AI error which was further reinforced through workshop exercises.

Our project was a pilot to develop methods and ways of working together to produce positive learning outcomes for librarians and their clients. We plan to further develop the project to make exhibits more mobile and learning materials virtually open and accessible, to enable further opportunities for librarians to learn about genAI so that they can better support exhibits, and to expand opportunities for visitors to further engage with the exhibit. For now, we have demonstrated how generative it could be for AI literacy initiatives to experiment with the concept of AI error and uncertainty at a local level if those initiatives are situated, speculative and participatory. There is much more work to be done across a much larger cohort, and our approach is but one among many working towards the goal of AI literacy for all.

Acknowledgements

We would like to thank workshop participants and exhibition hosts Nisa Alam, Alycia Bailey, Naomi Burns, Heather Davis, Abby Dawson, David Dolan, Ashley England, Kathryn Joss, Trish Hepworth, Fiona Ni, Kimberley Porteous, Nicole Powell, Anna Sin, Emma Slee, Emma Stockburn and Heidi Wills. Thanks to Jodi Schneider whose open access interview questions inspired our own, to Eva Harbridge and Tim Koskie for research assistance and Tony Jones for his fabrication work.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

We acknowledge the University of Technology Sydney and the Australian Research Council (FT240100011) for their generous funding of this work.

Notes on Contributors

Heather Ford is Professor in the School of Communication at the University of Technology Sydney (UTS) and an ARC (Australian Research Council) Future Fellow working on critical AI literacies. With a background working for non-profit technology companies in South Africa, Kenya, the United States and the United Kingdom, she now conducts research on how digital platforms shape what counts as knowledge and works with libraries, educational institutions and civil society organisations to improve digital and AI literacies for all. Heather completed her DPhil (PhD) at the Oxford Internet Institute at Oxford University and completed her Master's in Information Management and Systems (MIMS) at the University of California, Berkeley iSchool.

Andrew Burrell is a Senior Lecturer of Visual Communication, School of Design at the University of Technology Sydney. They are a practice-based researcher and educator exploring virtual and digitally mediated environments as a site for the construction, experience and exploration of memory as narrative. Their ongoing research investigates the relationship between imagined and remembered narratives and how the multi-layered biological and technological encoding of human subjectivity may be portrayed within and inform the design of virtual environments. Andrew's networked projects in virtual and augmented environments have received international recognition. Andrew uses creative practice to research and understand the complexities of emerging and speculative technologies, and is particularly interested in how these are implicated in more-than-human ecologies.

Monica Monin is a Lecturer of Visual Communication in the School of Design at the University of Technology Sydney. As a practice-based researcher, she uses creative practice as a vital way to research, understand and critically engage with emerging technologies. Her creative works are usually idiosyncratic, ongoing, dynamic processes which take the form of experimental infrastructures, installations, text or digital processes. Monin completed her PhD at the University of New South Wales where she developed process-oriented approaches of and for machine learning through a critical creative practice.

Bhuva Narayan is Associate Professor in the School of Communication, Faculty of Design and Society at the University of Technology Sydney (UTS). Bhuva was Head of Discipline for the Information and Knowledge Management Programme at UTS from 2013 to 2018, and Director of the Graduate

Research Programme at the Faculty of Arts and Social Sciences from 2019 to 2024 and is currently the Co-Director of the Graduate Research Programme at the new Faculty of Design and Society at UTS. Her research areas are in library and information science, specifically information behaviours, information avoidance, information access, user experience, human–computer interaction, personal information management, documentation studies, scholarly communication and open access. She is associate editor of the *Journal of the Australian Library and Information Association* (JALIA).

Suneel Jethani completed his PhD at the University of Melbourne and is Senior Lecturer of Digital and Social Media at the University of Technology Sydney. His research focuses on the politics of embodied biosensors and wearable technologies, critical data studies and ethics. His work has been published in journals including Continuum, Persona Studies, Communication, Politics & Culture, Cultural Studies, Body, Space & Technology and Conjunctions: Transdisciplinary Journal of Cultural Participation. He is the author of the monograph *The Politics and Possibilities of Self-Tracking Technology: Data, Bodies and Design (2021).*

ORCID

Heather Ford http://orcid.org/0000-0002-3500-9772
Andrew Burrell http://orcid.org/0000-0002-1690-7542
Monica Monin http://orcid.org/0000-0003-3316-6413
Bhuva Narayan http://orcid.org/0000-0001-8852-5589
Suneel Jethani http://orcid.org/0000-0003-2134-0904

References

Abdilla, A., Kelleher, M., Shaw, R., & Yunkaporta, T. (2021). *Out of the Black box: Indigenous protocols for AI*. Deakin University. Report. https://hdl.handle.net/10536/DRO/DU:30159239

AI, Algorithmic and Automation Incidents and Controversies (2025). *AIAAIC – AIAAIC repository: AI, algorithmic and automation incidents and controversies*. [online] www.aiaaic.org. https://www.aiaaic.org/aiaaic-repository

AI Incident Database (2025). *Welcome to the artificial intelligence incident database*. [online] incidentdatabase.ai. https://incidentdatabase.ai

American Library Association (2015). *Association of college and research libraries' framework for information literacy for higher education*. [online] Association of College & Research Libraries (ACRL). https://www.ala.org/acrl/standards/ilframework

Ananny, M. (2022). Seeing like an algorithmic error: What are algorithmic mistakes, why do they matter, how might they be public problem? *Yale Journal of Law & Technology*, *24*, 342–364. https://yjolt.org/seeing-algorithmic-error-what-are-algorithmic-mistakes-why-do-they-matter-how-might-they-be-public

Ananny, M. (2024). Making generative artificial intelligence a public problem. Seeing publics and sociotechnical problem-making in three scenes of AI failure. *Javnost – The Public*, *31*(1), 89–105. https://doi.org/10.1080/13183222.2024.2319000

Ananny, M., & Hudson, S. (2025). Oops? Interdisciplinary stories of sociotechnical error| oops? Sociotechnical errors as interdisciplinary stories of complex relations, shared consequences, and resilient hopes – introduction. *International Journal of Communication*, *19*(0), 2549-2553. https://ijoc.org/index.php/ijoc/article/view/24985

Australian Library and Information Association. (2025). The ALIA library and information services workforce framework. https://alia.org.au/common/Uploaded%20files/ALIA-Docs/2025/Revised_Framework_July_2025[7].pdf

Barassi, V. (2023). Special issue 5: Grappling with the generative AI revolution. *Harvard Data Science Review*, *Special Issue 5*, 1–6. https://doi.org/10.1162/99608f92.ad8ebbd4

Barocas, S., Crawford, K., Shapiro, A., & Wallach, H. (2017). The problem with bias: From allocative to representational harms in machine learning. *Special Interest Group for Computing,* *Information and Society (SIGCIS* 2017). http://meetings.sigcis.org/uploads/6/3/6/8/6368912/program.pdf

Barrabi, T. (2024). *Google moving to remove bizarre AI search results, like telling users to eat rocks* [Magazine]. New York Post, May 28, 2024. https://nypost.com/2024/05/28/business/google-moving-to-remove-bizarre-ai-search-results-like-telling-users-to-eat-rocks/

Bashir, N., Donti, P., Cuff, J., Sroka, S., Ilic, M., Sze, V., Delimitrou, C., & Olivetti, E. (2024). The climate and sustainability implications of generative AI. *An MIT Exploration of Generative AI*. https://mit-genai.pubpub.org/pub/8ulgrckc/release/2

Belot, H. (2023). Australian academics apologise for false AI-generated allegations against big four consultancy firms. *The Guardian*, 2 November 2023. https://www.theguardian.com/business/2023/nov/02/australian-academics-apologise-for-false-ai-generated-allegations-against-big-four-consultancy-firms

Carrick, D. & Kesteven, S. (2023). "Use with caution": how ChatGPT landed this US lawyer and his firm in hot water, *ABC News*, 24 June. https://www.abc.net.au/news/2023-06-24/us-lawyer-uses-chatgpt-to-research-case-with-embarrassing-result/102490068

Cizek, K., & Uricchio, W. (2022). *Collective wisdom: Co-creating media for equity and justice* (1st ed.). The MIT Press. https://doi.org/10.7551/mitpress/13394.001.0001

Cox, A. (2023). How artificial intelligence might change academic library work: Applying the competencies literature and the theory of the professions. *Journal of the Association for Information Science and Technology*, 74(3), 367-380. https://asistdl.onlinelibrary.wiley.com/doi/10.1002asi.24635

Cox, A. M., & Mazumdar, S. (2024). Defining artificial intelligence for librarians. *Journal of Librarianship and Information Science*, *56*(2), 330–340. https://doi.org/10.1177/09610006221142

Dao, D. (2023). *Awful AI.* [online] GitHub. https://github.com/daviddao/awful-ai

Department of Industry Science and Resources. (2025). *Developing a national AI capability plan.* Australian Department of Industry, Science and Resources. https://www.industry.gov.au/news/developing-national-ai-capability-plan

Dunne, A., & Raby, F. (2013). *Speculative everything: Design, fiction, and social dreaming* (1st ed.). The MIT Press.

Elish, M. C. (2019). Moral crumple zones: Cautionary tales in human-robot interaction. *Engaging Science, Technology, and Society*, *5*, 40–60. https://doi.org/10.17351/ests2019.260

Feffer, M., Martelaro, N., & Heidari, H. (2023, October). The AI Incident Database as an educational tool to raise awareness of AI harms: A classroom exploration of efficacy, limitations, & future improvements. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (pp. 1–11).

Ford, H. (2022). *Writing the revolution: Wikipedia and the survival of facts in the digital age.* MIT Press.

Freire, Paulo. (2000). *Pedagogy of the oppressed.* 30th Anniversary Edition. New York: Continuum. (Original work published 1970).

Giroux, H. A. (1988). Literacy and the pedagogy of voice and political empowerment. *Educational Theory*, *38*(1), 61–75. https://doi.org/10.1111/j.1741-5446.1988.00061.x

Hider, P., Rankin, C., Wakeling, S., Garner, J., & Jamali, H. R. (2023). Occupations and preoccupations of the Australian library profession: An analysis of job advertisements and professional literature. *Journal of the Australian Library and Information Association*, *72*(3), 225–250. https://doi.org/10.1080/24750158.2023.2233751

Kensing, F., & Blomberg, J. (1998). Participatory design: Issues and concerns. *Computer Supported Cooperative Work (CSCW)*, *7*(3), 167–185. https://doi.org/10.1023/A:1008689307411

King, J., Klyman, K., Capstick, E., Saade, T., & Hsieh, V. (2025). *User privacy and large language models: An analysis of frontier developers' privacy policies* (No. arXiv:2509.05382). arXiv. https://doi.org/10.48550/arXiv.2509.05382

Knight, S., McGrath, C., Viberg, O., & Pargman, T. C. (2025). Learning about AI ethics from cases: A scoping review of AI incident repositories and cases. *AI and Ethics*, 5(3), 2037–2053. https://doi.org/10.1007/s43681-024-00639-8

Leander, K. M., & Burriss, S. K. (2020). Critical literacy for a posthuman world: When people read, and become, with machines. *British Journal of Educational Technology*, *51*(4), 1262–1276. https://doi.org/10.1111/bjet.12924

Marres, N., Castelle, M., Gobbo, B., Poletti, C., & Tripp, J. (2024). AI as super-controversy: Eliciting AI and society controversies with an extended expert community in the UK. *Big Data & Society*, *11*(2), 1–18. https://doi.org/10.1177/20539517241255103

Mckie, I. A. S., & Narayan, B. (2019). Enhancing the academic library experience with chatbots: An exploration of research and implications for practice. *Journal of the Australian Library and Information Association*, *68*(3), 268–277. https://doi.org/10.1080/24750158.2019.1611694

Munn, L. (2024). The five tests: Designing and evaluating AI according to indigenous Māori principles. *AI & Society*, 39(4), 1673–1681. https://doi.org/10.1007/s00146-023-01636-x

Oddone, K., Garrison, K., & Gagen-Spriggs, K. (2024). Navigating generative AI: The teacher librarian's role in cultivating ethical and critical practices. *Journal of the Australian Library and Information Association*, *73*(1), 3–26. https://doi.org/10.1080/24750158.2023.2289093

Rothman, J. (2023). *Why the Godfather of A.I. Fears What He's Built.* [online] The New Yorker, 20 November 2023. https://www.newyorker.com/magazine/2023/11/20/geoffrey-hinton-profile-ai

Sanders, E. B. N., & Stappers, P. J. (2008). Co-creation and the new landscapes of design. *CoDesign*, *4*(1), 5–18. https://doi.org/10.1080/15710880701875068

Skibba, R. (2025, September 12). Are we offloading critical thinking to chatbots? *Undark Magazine*. https://undark.org/2025/09/12/critical-thinking-chatbots/

Suresh, H., & Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, https://doi.org/10.1145/3465416.3483305

Swan, D. (2024, February 12). ChapGPT lawsuit dropped by Australian mayor Brian Hood *Sydney Morning Herald*. https://www.smh.com.au/technology/australian-mayor-abandons-world-first-chatgpt-lawsuit-20240209-p5f3nf.html

Touretzky, D., Gardner-McCune, C., Martin, F., & Seehorn, D. (2019). Envisioning AI for K-12: What should every child know about AI?. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 9795–9799. https://doi.org/10.1609/aaai.v33i01.33019795

UNESCO et al. (2019). Beijing consensus on artificial intelligence and education. https://unesdoc.unesco.org/ark:/48223/pf0000368303

van der Riet, M. (2008). Participatory research and the philosophy of social science. *Qualitative Inquiry*, *14*(4), 546–565. https://doi.org/10.1177/1077800408314350

Velander, J., Otero, N., & Milrad, M. (2024). *What is critical (about) AI literacy? Exploring conceptualizations present in AI literacy discourse.* In A. Buch, Y. Lindberg, & T. Cerratto Pargman (Eds.), *Framing futures in postdigital education* (pp. 139–160). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-58622-4_8

Vincent, J. (2019). *'Godfathers of AI' honored with turing award, the nobel prize of computing*. The Verge. https://www.theverge.com/2019/3/27/18280665/ai-godfathers-turing-award-2018-yoshua-bengio-geoffrey-hinton-yann-lecun

Yerbury, H., & Henninger, M. (2024). *Journal of the Australian Library and Information Association*, *73*(3), 362–379. https://doi.org/10.1080/24750158.2024.2362973

Zhang, P., Zeng, G., Wang, T., & Lu, W. (2024). *TinyLlama: An open-source small language model* (No. arXiv:2401.02385). arXiv. https://doi.org/10.48550/arXiv.2401.02385

Appendix

Interview Guide

(A) Introductory questions.
  (1) Can you tell me the story of how you came to work as a librarian?
  (2) Can you describe a typical day's work for you at the library?
  (3) How do you usually support queries from library clients about how to find information on particular topics? What tools, resources and processes do you use?

(B) Supporting information discovery.

(1) I would like you to recall a specific incident where you were gathering information for a client on a particular topic.
 (a) [Prompt to recall/describe]
 (b) For what purpose were you gathering this information?
 (c) Tell me a little bit about what this process looks like for you. [Have them tell you how they go through this process from start to finish]

(2) Can you talk about what that experience was like?

(3) Describe to me how you went about finding that information.

(4) What types of information did you look for? (primary/secondary/tertiary sources, scientific articles, experiential stories, statistics)

(5) How did you decide which type of information would be most beneficial for your needs?

(6) Was there anything you didn't find credible? Why? How did you know it wasn't credible?

(7) How did you decide what information was relevant versus irrelevant? (i.e. How did you decide which information is usable?)

(8) How did you know a source was credible? What indicates that to you? How did you know?

(9) What was the most challenging part of this process? What was the easiest?

(10) Did you seek assistance and/or collaborate with others? If so, what did that process look like?

(11) How did you see your role in relation to information? How did you see yourself?

(C) Generative AI

(1) Have you used ChatGPT before?

(2) What is your level of confidence in being able to support clients' use of chatGPT?

(3) What do you think generative AI is useful for?

(4) What queries do you receive from clients about the use of generative AI tools like ChatGPT?

(5) What resources does the library have to support you and your clients' use of generative AI tools such as ChatGPT (if any)?

(6) What are the challenges in continuing to support your clients' search and discovery of information in the face of generative AI tools like ChatGPT?